OPO OPHTHALMIC & PHYSIOLOGICAL OPTICS
THE JOURNAL OF THE COLLEGE OF OPTOMETRISTS

Ophthalmic & Physiological Optics ISSN 0275-5408

Recognition acuity in children measured using The Auckland Optotypes

Lisa M Hamm[1,2], Nicola S Anstice[1,2,3], Joanna M Black[1,2] and Steven C Dakin[1,2,4]

[1]School of Optometry and Vision Science, The University of Auckland, Auckland, New Zealand, [2]New Zealand National Eye Centre, Auckland, New Zealand, [3]School of Optometry and Vision Science, University of Canberra, Canberra, Australia, and [4]UCL Institute of Ophthalmology, University College London, London, UK

Citation information: Hamm LM, Anstice NS, Black JM, & Dakin SC. Recognition acuity in children measured using The Auckland Optotypes. *Ophthalmic Physiol Opt* 2018; 38: 596–608. https://doi.org/10.1111/opo.12590

Keywords: acuity, agreement, children, optotypes, psychophysics, screening

Correspondence: Lisa Hamm
E-mail address: l.hamm @auckland.ac.nz

Received: 24 April 2018; Accepted: 2 November 2018

Abstract

Purpose: Sloan letters displayed by the Electronic Visual Acuity (EVA) system are the gold standard for recognition acuity measurement in research settings. However, letters are not always appropriate for children. The Auckland Optotypes (TAO) are a new, open-access set of 10 pictograms available in regular and vanishing formats. We sought to assess feasibility of using both formats of TAO for measuring visual acuity (VA) in children using a Bayesian adaptive staircase, in a community setting.

Methods: We tested 121 children (5–12 years old) with both formats of TAO, a handheld flipchart vision screener (Parr vision test), as well as the gold standard EVA. We measured feasibility of the three comparison tests in three ways. First, using limits of agreement (LoA) with EVA, second, calculating area under the receiver operating characteristic curve (AUC), and finally, investigating trial-by-trial responses.

Results: Agreement between tests was within test-retest reliability of EVA measures ($LoA_{TAOregular} = \pm 0.14$, $LoA_{TAOvanishing} = \pm 0.15$, $LoA_{Parr} = \pm 0.16$ logMAR). TAO tests were highly effective at identifying children with vision impairment ($AUC_{TAOregular} = 0.96$, $AUC_{TAOvanishing} = 0.95$), whereas Parr was less effective ($AUC_{Parr} = 0.82$). In 5–6 year old children there was an enhanced advantage of TAO ($AUC_{TAOregular} = 0.97$, $AUC_{TAOvanishing} = 0.98$) over Parr ($AUC_{Parr} = 0.75$). Although each child completed 16 trials, approximately 10 trials were sufficient to achieve excellent LoA, and six trials sufficient for accurate screening.

Conclusion: Threshold VA assessment and vision screening are feasible using both vanishing and regular formats of TAO.

Introduction

The Electronic Visual Acuity (EVA) testing system[1–3] is a valuable tool for measuring the visual acuity (VA) of children within a research setting (e.g. for clinical trials, as used recently by Guo *et al.*[4]). The EVA uses the ETDRS set of 10 Sloan letters for children 7 years and older, and the four item HOTV Sloan letter set for children younger than 7 years.[5] For an observer unfamiliar with Roman letters, having fewer alternatives can aid identification, but also increases the possibility of correct guessing, which can lead to over-estimation of performance.[6] An alternative approach is to use pictogram optotypes with more than four options.

We have recently described a new set of 10 pictogram optotypes (The Auckland Optotypes, or TAO)[7] designed to be used across cultures, and in either regular (black presentation on a white background with surrounding interaction/crowding bars) or vanishing (split black and white strokes on a grey background with no crowding bars[8]) formats. TAO has two key advantages over other optotype sets. First, TAO are more accessible for participants who do not know Roman letters. Second, each shape elicits an

acuity estimate that is more similar to others within the set compared to Sloan letters.[7]

Vanishing optotypes are unique in that the split black and white strokes on a grey background mean that when this luminance modulation cannot be resolved the optotype appears to vanish into the grey background.[8] This contrasts with traditional black letters on a white background, which retain low spatial frequency information after the stroke is unresolvable. The Cardiff Acuity Test[9] uses this vanishing format for testing children's acuity within a forced choice preferential looking task, while the TAO requires children to name the optotype from the set of 10 available options. Vanishing variants were included in this project as they have superior test-retest reliability[10], and may be more sensitive to measuring subtle changes in acuity arising from some conditions (for example, macular degeneration[11]).

Regardless of the optotype set or format being presented, the choice of what size to present the target is central to measurement of VA. This includes both the protocol through which a threshold is estimated, as well as a definition of what constitutes 'normal' performance. In terms of protocol, the size-progressions used clinically remain largely based on standards established with physical charts.[12–14] Indeed, digital tests often use a similar presentation protocol to charts.[15–17] Standardised digital tests for children have favoured a chart-style format, but opting for single optotype presentation and phases of testing (engagement, reinforcement, etc[1,2]). In a research setting, however, Bayesian adaptive staircases (such as QUEST[18] and ZEST[19]) are popular (for an overview see Klein[20]). These algorithms take full advantage of single item presentation through use of variable step sizes which, together with their use of all observer responses, supports more efficient threshold-estimation.[21] We were interested in whether a Bayesian strategy might be feasible/useful for VA testing of children.

In terms of what constitutes 'normal' performance, different sets of symbols elicit different VA results, even if the dimensions of the optotypes (stroke width and bounding box) are matched.[16] In fact, normal performance depends on the subset of symbols within the set (consider the difference in VA elicited from the ETDRS and HOTV subsets of Sloan letters).[15,16] A choice needs to be made, therefore, about whether a recognition task is intended to capture the minimum angle of resolution (the stroke width facilitating recognition), the overall size of a symbol which can be recognised (bounding box size), or whether symbols should be scaled so that performance falls in line with performance measured with other sets. Although the latter case is the least principled, it is the only case in which acuity results between optotype sets can be compared (offering substantial clinical utility), and has been the choice of most picture optotype sets.[22,23]

This study was conducted in New Zealand (NZ) where the Parr chart, a modified version of the Sheridan Gardiner test, is used for pre-school vision screening in children 4–5 years of age. This test consists of seven Sloan font letters (A, H, O, T, U, V, X), presented within a 15-page flipbook where each page shows one optotype surrounded by crowding bars. Given its simplicity and importance within NZ, we included this test as an additional comparison to the regular and vanishing TAO tests. We have previously reported acuity estimates in adults for the TAO set including inter-optotype reliability.[7] Building on this work, here we explore the feasibility of the newly developed TAO set (in regular and vanishing formats) for testing recognition acuity of children in a community setting.

We used a QUEST staircase procedure to set stimulus presentation size. We then used Bland Altman analysis[24] to estimate 95% limits of agreement (LoA) between TAO and EVA, and compared this to established test-retest performance of the EVA testing system.[1,2] We use all three alternatives for calculating logMAR (stroke, bounding box and subjective equivalence) and evaluate mean agreement with EVA for each alternative. In addition to quantifying agreement with a standard test, we consider a test's capacity to identify children with a visual problem. This is typically done by assessing sensitivity and specificity from receiver operating characteristic (ROC) curves after separating participants into those with from those without a visual problem. To further explore feasibility of the TAO tablet test, we investigated which optotypes are correct most often as well as how many trials are needed to achieve acceptable LoA with EVA[1,2] and to maintain diagnostic accuracy.

Methods

Participants

The project was approved by The University of Auckland Human Ethics Committee, and our protocol complied with the tenets of the Declaration of Helsinki. Parents provided written, informed consent and children provided written assent. One hundred and twenty-one children were recruited from three low socioeconomic status, culturally diverse schools within the Auckland region. We included children aged 5–12 years because we were equally interested in threshold estimation (relevant for research and clinical application throughout childhood) as well as pass/fail screening status (most relevant to preschool screening). All schools taught in English. Parents reported the child's ethnicity and when self-identification included mixed backgrounds, we used the first ethnicity reported. Ethnicity was grouped based on Level 1 classifications from Statistics NZ.[a] Age, gender and ethnicity characteristics for our cohort are summarised in *Figure 1*.

[a]http://archive.stats.govt.nz/methods/classifications-and-standards/classification-related-stats-standards/ethnicity.aspx (accessed 24 April, 2018).

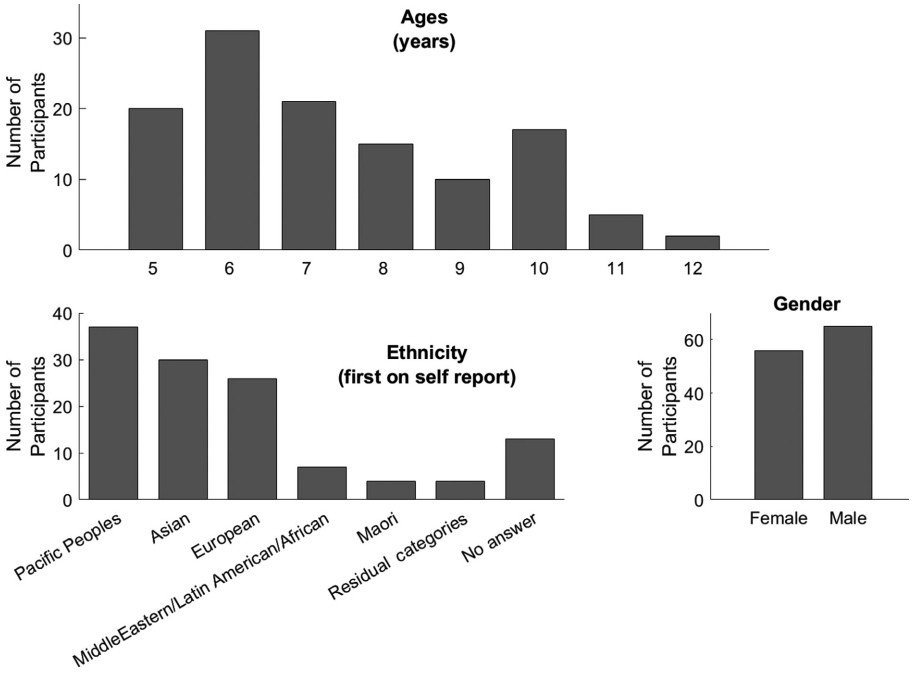

Figure 1. Demographic information.

Visual assessment

Testing was conducted where it was most convenient for the participant's family (who were given the choice of the school or The University of Auckland Eye Clinic). For clinic appointments, parents accompanied their child, whereas for testing at the school, parents were given the option to attend or receive a post-assessment phone call to discuss outcomes. Two trained researchers conducted the acuity testing: one researcher ran the EVA and Parr, and the other ran both variants of TAO test on a tablet computer. Children were tested in groups of two, allowing test order to be pseudo-randomised between EVA and TAO, i.e. half the children completed the TAO task first while the other completed EVA acuity measures first. We tested right eye, then left eye using occluding glasses. If a child had refractive correction we tested uncorrected vision first, then corrected. We henceforth report uncorrected VA.

Acuity-specific protocols

Prior to testing, we ensured each child could identify the symbols in a given set by matching on a key card. For EVA and Parr, the child was shown each target in paper format and either named or pointed to the corresponding item on their key card. For the tablet tests, we formalised this process by showing children short animations[b] in which a cartoon of the optotype turned into the black symbol on a white screen. The optotype was displayed within a 4 cm by 4 cm box, such that the minimum angle of resolution (stroke width) subtended approximately 10 min of arc (similar to 1.0 logMAR, or Snellen equivalent 6/60). A single frame from each animation is shown in *Figure 2* with the corresponding optotype. Children were asked to name or match the symbol after each animation. If a symbol was identified incorrectly, the symbol was displayed again at the end of the set of animations. If incorrect a second time, the child was deemed untestable. This phase served a similar function to the 'pre-test' described by Holmes[1]. We did not provide formal feedback during any task, but children were provided with continued encouragement, and praise for providing answers, whether correct or incorrect. Assessors were privy to stimulus level on the EVA and Parr tests, but not the TAO tests. The TAO tests had a re-randomise option, whereby the assessor could display a second optotype of the same size if the child lost concentration; this feature was not available on the EVA or Parr tests.

Every regular format optotype test (TAO, EVA and Parr) utilised single optotype presentation with crowding bars. Vanishing optotypes (TAO only) were not crowded. For TAO and Parr, bars were positioned half an optotype width (edge to edge distance) away from the symbol. For EVA, crowding bars were positioned one optotype width from the symbol. Testing distance, stimulus size progression protocol and termination criteria varied between

[b]https://github.com/dakinlab/OpenOptotypes (accessed 24 April, 2018).

Figure 2. Frame from animation with corresponding optotype.

Table 1. Test overview

	Test distance	Protocol	Termination	Testable range (logMAR)	Optotypes
The Auckland Optotypes	1.5 m	QUEST adaptive (16 trials)	Trials complete	−0.3 to 1.5	
Parr	4 m	3 optotypes at each level (5 levels)	2 incorrect at one level	0.0 to 0.7 (converted from Snellen)	**A H O T U V X**
Electronic Visual Acuity	3 m	Custom adaptive algorithm	Algorithm complete	−0.2 to 1.6	**C D H K N O R S V Z (H O T V)**

tests, as summarised in *Table 1*. We used the standardised protocols for both EVA[1–3] and Parr[c] (including their use of logMAR approximations for Snellen stimulus size). For TAO tablet tests, we used a similar protocol to a previous version[25] based on a Bayesian adaptive staircase (QUEST)[18], programmed in MATLAB (www.mathworks.com) using Psychtoolbox (www.psychtoolbox.org,[26,27]). We set the expected acuity threshold to 0.0 logMAR, with a standard deviation of 0.3 logMAR, estimated lapse rate to be 1% and the guessing rate to 10% (to reflect the 10 alternatives possible). Rather than starting at threshold, we forced the first presentation to a larger and therefore easier 0.3 logMAR to increase motivation. If the first presentation was correct, the second presentation was set to 0.15 logMAR, and thereafter (or if the first presentation was incorrect) progression was left to the QUEST algorithm. We presented 16 trials per staircase. QUEST generated a threshold estimating the stimulus size supporting 75% correct identification. The test was run on a Microsoft Surface Pro 3 tablet computer (www.microsoft.com). The built-in LCD display (2160 × 1440 pixels, subtending 9.6° by 6.5°) was gamma corrected (white, ~300 cd m^{-2}, black 1 cd m^{-2} and grey 150 cd m^{-2}). Testing was conducted at a viewing distance of 1.5 m, based on recommendations for screening preschool children.[5]

[c]https://www.health.govt.nz/system/files/documents/publications/national-vision-hearing-screening-protocols-v3.pdf

Comprehensive eye examinations

Complete paediatric eye examinations were conducted by eye care professionals experienced in providing care to children. The eye examination included unilateral and alternating cover test performed at distance (≥3 m) and near (40 cm), near point of convergence, stereoacuity with the Randot preschool test, ocular motility testing, cycloplegic refraction and ocular health assessment. Retinoscopy was performed a minimum of 35 min after installation of one drop of 1% cyclopentolate. Dilation was considered complete if the pupil size was >6 mm and the pupillary reflex was absent 40 min after instillation. One further drop of cyclopentolate 1% was added, at the discretion of the clinician, after 40 min if required. The fundus and ocular media were assessed using direct and binocular indirect ophthalmoscopy. Only participants whose families agreed to have their child undergo cycloplegic refraction were included in the study, however, two children refused cycloplegia on the day of testing and their acuity data is included in our results.

Analysis

When converting between pixel size and logMAR for Sloan letters, the height of the overall optotype size (bounding box) is divided by 5, such that the minimum angle of resolution (MAR) is the stroke width. The logMAR system has been standardised such that a Sloan letter which subtends 5 min (and therefore a stroke that

subtends 1 min) has a common logarithm of the MAR (stroke) equal to 0. However, for optotypes with less of the bounding box comprised of the stroke (such as Lea symbols), direct use of stroke width as the MAR *overestimates* acuity in relation to Sloan letters,[28] and direct use of the bounding box *underestimates* acuity compared to Sloan letters.[29] If the goal is to have results equitable to Sloan VA tests, a modified divisor is often required to estimate logMAR (summary provided by Bailey and Lovie-Kitchin[15]). Although a full study would need to be done to establish appropriate performance-based scaling factors, prior to the current project we derived estimates from published accounts of scaling for other picture sets[28] and previous work with TAO symbols [comparing TAO to Landolt Cs in adults,[7] TAO to Lea symbols in children[25] and some pilot data comparing TAO to EVA results in children (Hamm, Anstice & Dakin, unpublished)]. This resulted in estimates of 0.0 logMAR equivalent for bounding box sizes at 7.6 and 12.6 min for regular and vanishing optotypes, respectively (summarised in *Figure 3*). Note that this scaling influences mean differences (bias) between EVA and TAO tests, but not the LoA between tests. Likewise, it influences acuity cut-offs for sensitivity and specificity, but not the area under the ROC curves. We will focus on measures independent of scaling, however, we report mean difference between EVA and TAO for each method of calculating logMAR, as well as reporting the

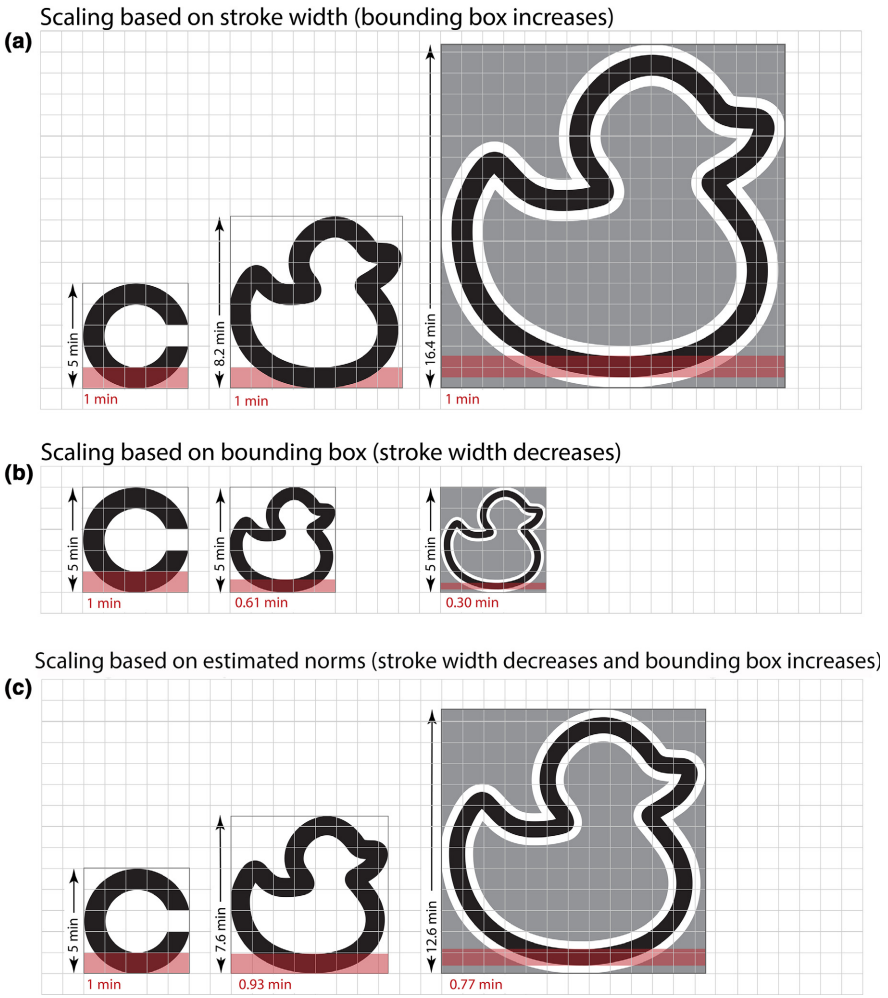

Figure 3. Scaling of The Auckland Optotypes (TAO). At the 0.0 logMAR line, Sloan letters (such as the **C** shown here) are displayed at 5 min of arc, such that the stroke width (minimum angle of resolution, or MAR) is 1 min. For TAO, which has a different bounding box to stroke ratio, matching stoke width (Panel a) causes an overestimation of acuity and matching bounding box (Panel b) causes TAO to underestimate acuity. If the goal is to have a new test elicit similar logMAR values, scaling needs to be based on normative data rather than bounding box or stroke width. Although we made estimates from previous projects and pilot data (Panel c), development of robust normative scaling factors requires more research.

scaling factor required in this study for the mean difference to be 0.0 logMAR.

Although QUEST was responsible for setting stimulus size during the psychophysical testing protocol using TAO, we re-fit raw trial-by-trial stimulus-response data with psychometric functions. We used the Palamedes tool box[30] to do this, and set the acuity range from −0.3 to 1.5 logMAR with the slope from 3.3 to 50. Lapse and guess rates were fixed at 1% and 10%, respectively. We used the PAL_PFML_Fit function and estimated threshold from the best-fitting cumulative normal distribution.

For Bland Altman analysis, we used the eye with poorer vision based on EVA scores. In cases of equal VA, we used data from the right eye. Test-retest reliability for the EVA was estimated from two previous studies[1,2] by calculating 95% LoA. These values were ±0.21 logMAR and ±0.19 logMAR respectively. Given the use of EVA as a gold standard, we reasoned acceptable LoA with EVA would be near ±0.2 logMAR.

Each participant was assigned a status (visually normal or visually impaired) using the American Association of Pediatric Ophthalmology and Strabismus (AAPOS) criteria[31,32], (for details see Column 2 of *Table 2*). Based on these binary classifications, ROC curves were generated, and area under the curve (AUC) was used as a measure of test utility for screening (from which sensitivity and specificity of each test can be calculated at various cut-offs). An area of 1 under the ROC curve indicates perfect sensitivity and specificity, while 0.5 indicates the test has no predictive power (the test result provides only chance levels of distinguishing a 'patient' from a 'control').

In addition to these two key outcome measures (LoA and AUC), we further explored TAO tests results. We inquired about lapse rates, duration of task, bias in optotype responses, and impact of number of trials. We used repeated measures ANOVA and correlations as part of these additional investigations.

Results

Visual outcomes of children

All 121 children were deemed testable on TAO and EVA. One child did not have a measurable acuity on the Parr test, as he was unable to achieve two correct responses on the 0.7 logMAR level, which is the largest optotype (his VA was measured at approximately 1.0 logMAR using EVA). Criteria for visual impairment as well as the number of children meeting each criterion are presented in *Table 2*. Eleven participants out of 121 (9%) failed screening criteria. Five and six year olds accounted for less than half the participants (51 of 121), but over half of the children (6) identified with a vision problem. Two children had a vision problem without associated unaided acuity loss on the gold standard acuity test; one child had hyperopia and the other had manifest strabismus.

Agreement with EVA

Figure 4 shows the results of the TAO tests in both regular and vanishing formats as well as the Parr test, each compared to the EVA results in the form of Bland Altman plots.[24] Overall, mean acuity threshold estimates from TAO tests and EVA were very similar; with only approximately one letter (0.03 logMAR) difference between the EVA result and the result from both the regular and vanishing formats of the TAO test. In other words, with current use of 7.6 (regular) and 12.6 (vanishing) as divisors (scaling factor) in the logMAR calculation resulted in EVA yielding slightly better acuity (lower logMAR values) than TAO.

The differences between VA results were normally distributed for both TAO tests, but not Parr. Bland Altman LoA for the weaker eye were less than ±0.2 logMAR for all tests, the approximate test re-test reliability for EVA in children up to 7 years old[1,2]. The

Table 2. Number of participants who failed based on diagnostic criteria.

Diagnosis	Criteria	Number failed
Refractive error		**3** (2)
Myopia	More than 1.50 D	**0**
Hyperopia	More than 3.50 D	**1** (1)
Astigmatism	More than1.50 DC	**1** This child also failed visual acuity
Anisometropia	More than 1.50 D interocular difference (spherical equivalent)	**1** (1) This child also failed visual acuity
Strabismus	>8 manifest tropia	**1**
Unexplained acuity loss	Worse than or equal to 0.2 logMAR (Electronic Visual Acuity) and no other positive findings	**7** (4)
Total		**11** (6)

Total number in bold, and number of 5 and 6 year olds in parenthesis.

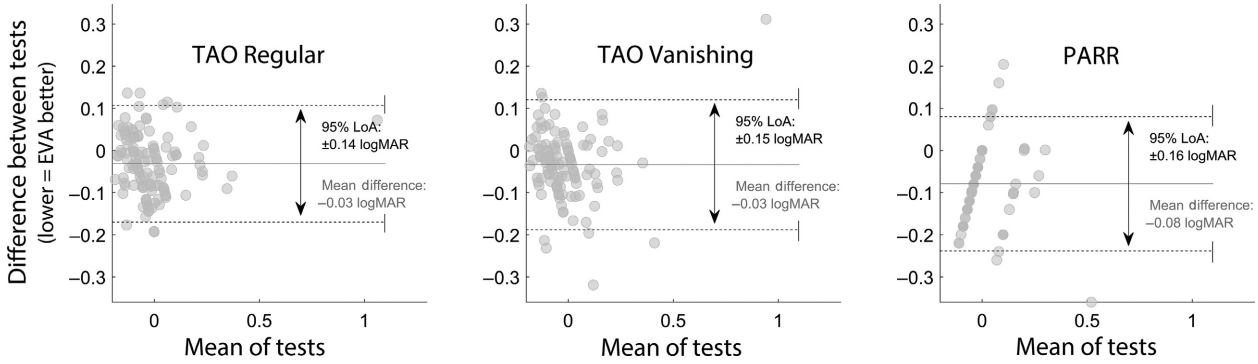

Figure 4. Bland Altman results for the weaker eye. Each test is a comparison of the named test with the gold standard Electronic Visual Acuity (EVA). Data points are partially transparent to allow viewing of overlapping data points. The solid horizontal grey line represents mean difference, dashed horizontal lines represent 95% limits of agreement (± 95% confidence intervals – solid vertical lines).

95% LoA for the regular version of TAO was ±0.14 logMAR, for the vanishing version was ±0.15 logMAR and for the Parr test was ±0.16 logMAR.

Influence of scaling factors on mean difference

Although we started with an estimate for scaling, we also calculated mean difference using the two other strategies outlined in *Figure 3*. Using the exact stroke width (1/8.2 of the bounding box) caused estimates for regular optotypes to overlap exactly with mean EVA results (0.0 logMAR difference). However, for vanishing optotypes, stroke width (1/16.4 of the bounding box) meant VA estimates became better than EVA (by 0.08 logMAR). Alternatively, if MAR estimates were based on simply dividing the bounding box by 5 (as is the case for Sloan letters), the estimates are less consistent with VA measures from the EVA test. Using this method of estimating logMAR, $TAO_{regular}$ had a mean difference of −0.22 logMAR and vanishing −0.44 logMAR, such that EVA scores were much better than TAO. For the TAO vanishing set to be perfectly aligned with EVA for this cohort, the divisor needed to be 13.5.

ROC curves

The regular and vanishing variants of the tablet TAO test were good at identifying children with a vision disorder based upon the AAPOS criteria (AUC = 0.96 and 0.95 respectively, *Figure 5*). Both performed better than the Parr test (AUC = 0.82). Sensitivity and specificity can be calculated for specific cut offs (*Figure 5* inset shows examples), but this is directly related to how the optotypes are scaled, which can be adjusted. Note the slight overestimation of acuity by TAO is associated with a lower (0.1 logMAR) cut off VA with the current scaling.

Analysis of children aged 5 and 6 years

Since we used a different EVA protocol (HOTV letters only) for children younger than 7 years, we also ran an adjunct analysis on only this subset of participants. The results are compiled into a single figure below (*Figure 6*). Limits of agreement were similar to the whole cohort (±0.13 to ±0.16 logMAR). Within these 51 participants, six had visual problems as defined by the AAPOS criteria. As with the full cohort, each tablet TAO test was more effective than the Parr test in terms of the AUC ($AUC_{TAOregular}$ = 0.97, $AUC_{TAOvanishing}$ = 0.98, AUC_{Parr} = 0.75). Note the difference in AUC between the tablet tests and the Parr test is larger in the subset of younger children than for the whole cohort.

Further analysis of tablet-based tests

Lapses

Lapses are errors made when the judgement should be easy, typically due to factors such as distraction. By counting trials on which children were shown an optotype well above their threshold and answered incorrectly, we can estimate a lapse rate for the cohort. We set the criterion for 'well above' at 0.1 logMAR above TAO test threshold for each staircase. For regular optotypes, the lapse rate was 0.67% of trials, and for vanishing it was 0.83%; both just under the 1% used for QUEST and the Palamedes refits, suggesting it is an appropriate estimate for this cohort and protocol used.

Duration

TAO test duration (for a single staircase) averaged 86 ± 30 s for the whole cohort and 90 ± 31 s for children 5 and 6 years old. *Figure 7* displays test duration (± standard deviation) left to right in order of task completion for the whole cohort. Repeated measures ANOVA results show that task duration shortened with experience ($F_{(3)}$ = 15.4,

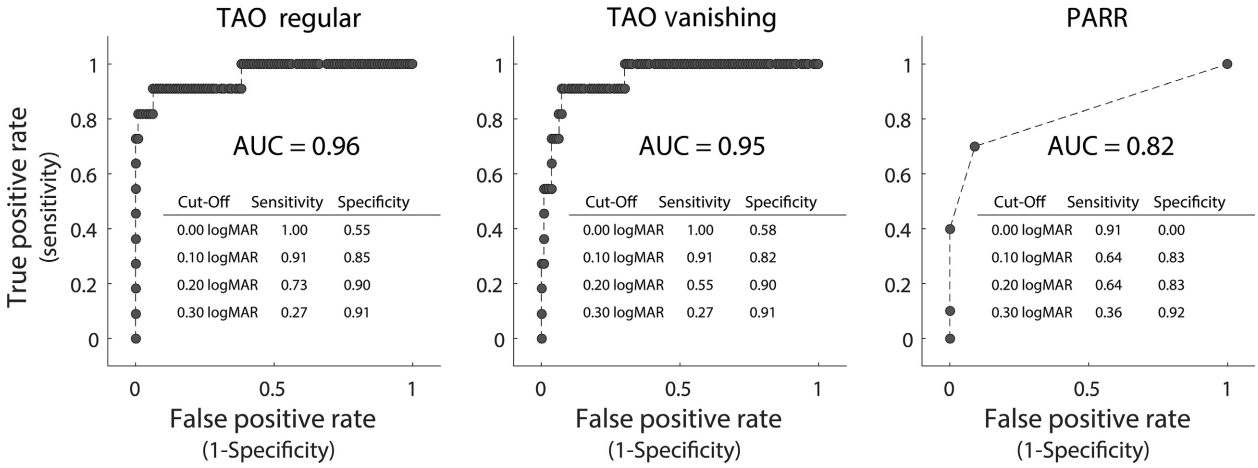

Figure 5. Receiver Operating Characteristic (ROC) curves. ROC curves for children's acuity tests for detecting failure of one of more of the AAPOS screening criteria.[31]

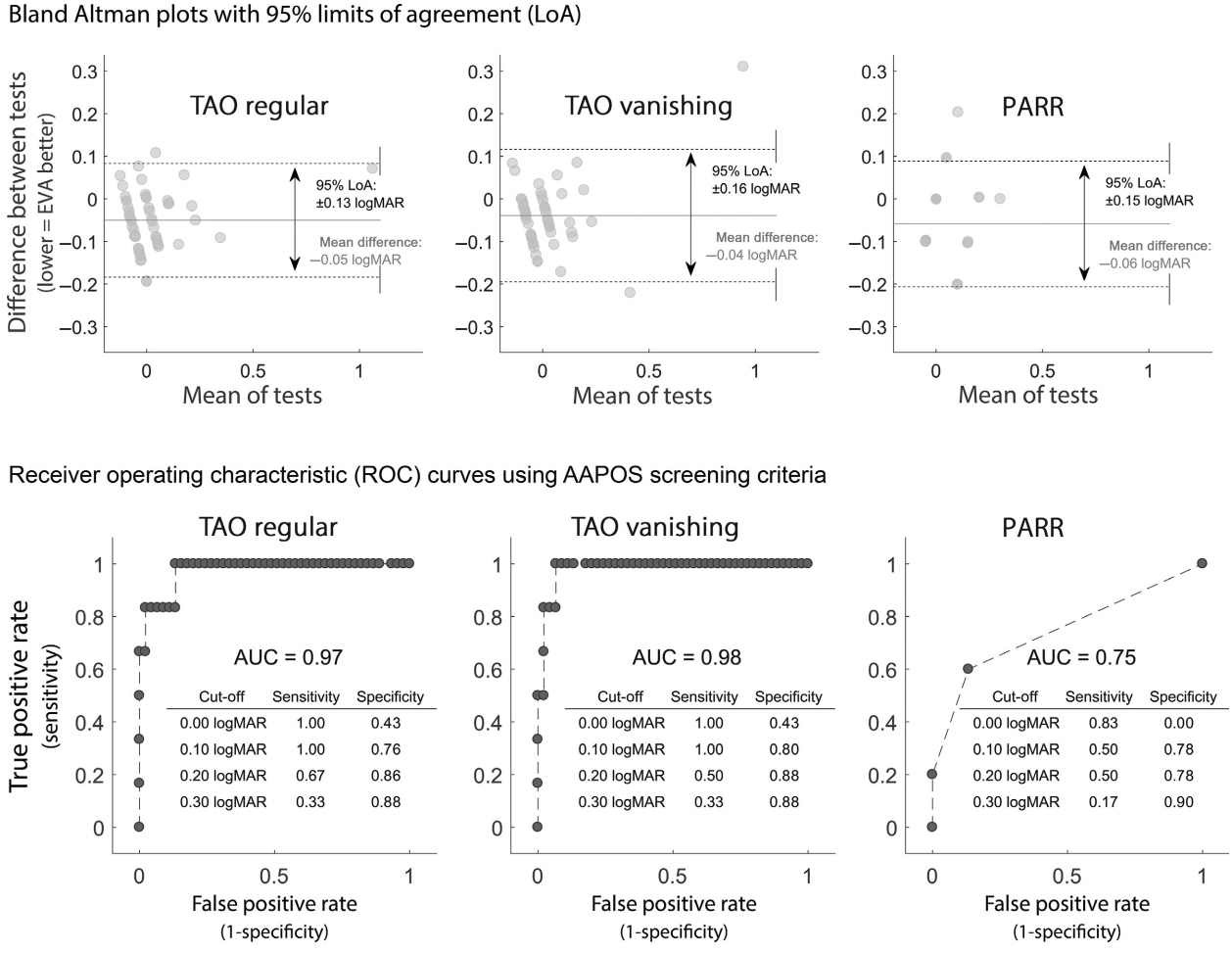

Figure 6. Results from subset of participants aged 5 or 6 years.

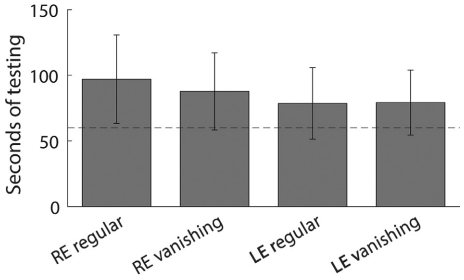

Figure 7. Test duration ± standard deviation. One minute is depicted with a dotted line for reference.

$p < 0.001$). Specifically, Bonferroni corrected post hoc analysis confirms that the RE regular (first test) took longer than both LE regular (third test) and vanishing (fourth test).

Bias

If certain optotypes were either easier to identify than others, or if there was a group level bias in favour of some shapes, we would expect the identification rate for some optotypes to be higher than for others. To look at this, we pooled trials across all participants and eyes (121 participants × two eyes, 16 trials each: 3872 total trials), separately for regular and vanishing stimuli, and looked at correct trials within each set. Across all trials, we expect the proportion correct to be approximately 0.10, assuming there is not differential difficulty or bias. Among correct trials for regular format optotypes, the proportion that each optotype contributed to a correct response ranged from 0.06 (178/2852) to 0.14 (390/2852). For vanishing optotypes the range was similar (0.07 to 0.13).

To disentangle the contribution of bias (some shapes being preferred) from difficulty (some shapes being easier to identify) we calculated ranked individual optotype difficulty from a previous project using TAO in adults[7]. Difficulty for regular and vanishing optotypes were not correlated with one another in this adult data set. We could therefore ask whether variation in children's responses were better explained by difficulty (assessed by correlation with ranked, individually assessed thresholds in the adult data) or by preference or 'bias' towards some shapes over others (assessed by correlation between proportional correct responses for individual regular and vanishing optotypes). Although both account for some variance, bias was a better predictor of correct responses. From the scatter plots in *Figure 8*, it is clear that the and were less popular responses than the , , ♡ and , particularly for younger children. Neither gender nor ethnicity appeared to change the pattern of correct responses (each correlation – male vs female, or any combination of ethnicities – showed at least $p < 0.05$ and $R^2 > 0.75$).

Trials

We used Palamedes refitting to examine how a reduction in the number of trials would influence outcomes. For the number of trials ranging from 1 to 16, we calculated the theoretical capacity for effective screening with AUC for ROC curves as well as theoretical LoA with EVA. The inset in each subplot in *Figure 9* shows how we derived a single data point in each subplot (in each case the inset shows the results at 16 trials). Row 1 is the impact of trial number on AUC results, using the AAPOS screening criteria and Row 2 shows LoA with EVA.

It is clear from *Figure 9* that additional trials yield diminishing returns in terms of identifying children with vision disorders. As such, the data were fit with exponential functions (orange lines). The 'knee' of each function was established by adding a straight line from the result at Trial 1 to the result at Trial 16, and finding the trial corresponding to the maximum perpendicular distance between the straight line and the fitted curve. We highlight this point (where additional trials produce diminishing returns) with an orange arrow and the corresponding (rounded) trial number.

Discussion

Results from the new TAO tablet test (using regular and vanishing optotypes) showed good agreement with the standard EVA test. Indeed, all test results were within test-retest scores for EVA,[1,2] and there was little difference in LoA across the three comparison tests ($LoA_{TAOregular} = \pm 0.14$, $LoA_{TAOvanishing} = \pm 0.15$, $LoA_{Parr} = \pm 0.16$ logMAR). Note that our children were older (5–12, rather than 3–7 years), and had less visual anomalies (as they were recruited from schools rather than clinics) than the participants in the two EVA test-retest studies.[1,2] Additionally, all children were tested by the same research team rather than at different sites, as was the case for the other published EVA studies.[1,2] The excellent agreement in our study therefore may be partially attributable to our cohort and protocol. Other authors have shown similar agreement between tests; for example, even in very young children (3–6 years), Moganeswari, *et al.*[33] reported agreement between Lea and HOTV charts of ±0.12 logMAR. Others have found poorer agreement; a new digital test showed 95% of participants had results within ±0.27 logMAR of a Snellen chart.[34] Additionally, beyond eight trials, we saw diminishing returns in terms of agreement with EVA, with 10 trials for regular and vanishing achieving LoA of ±0.15 logMAR. This suggests the TAO test could be shorted beyond the 1.5 min it was taking children per test, without compromising accuracy.

In clinical terms, accuracy is dependent on the actual acuity outcome rather than just the agreement. Such perceptual

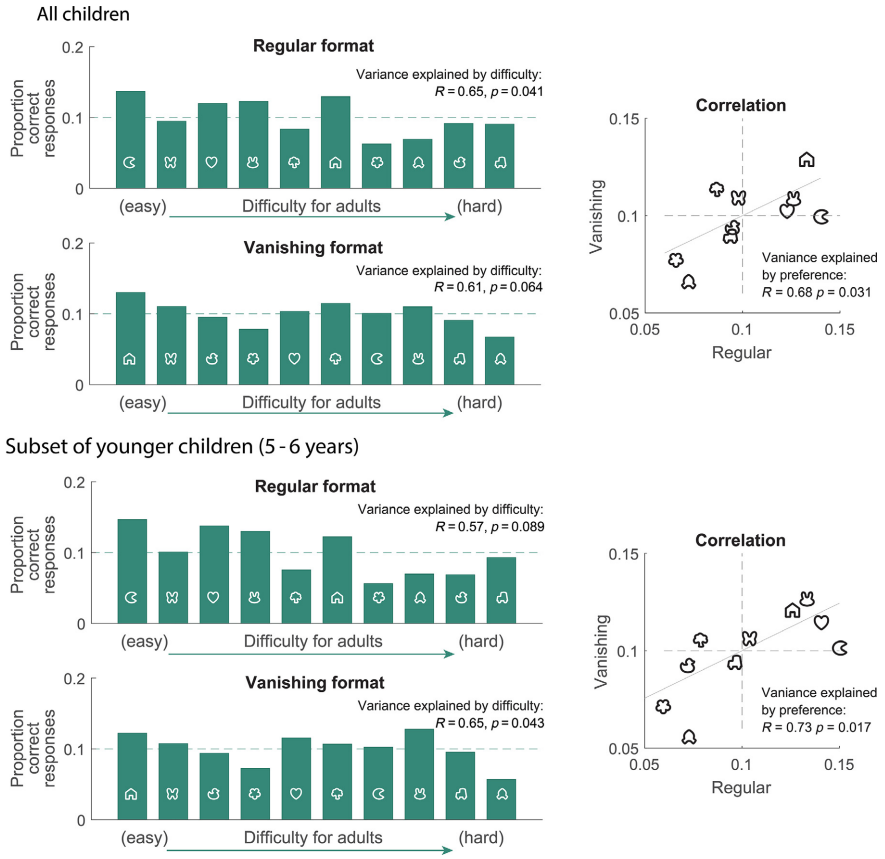

Figure 8. Correct responses by optotype. Proportion of correct responses is plotted for each optotype, with chance performance represented with a dotted line. Ranked difficulty estimates are from a recent study in adults which measured thresholds for individual optotypes.[7]

equivalence between optotype sets often requires a scaling factor. Here we trialled scaling factors (for regular and vanishing formats) derived from the description of Lea symbols by Cyert *et al.*[28] adjusted based on (unpublished) pilot data and a previous work investigating use of TAO regular and vanishing optotypes in children,[25] and adults.[7] Using these scaling factors, measured VA was slightly better (by approximately one letter) than EVA results. Interestingly, acuity results were more aligned with EVA when TAO regular optotypes were unscaled, and when the vanishing optotypes scaling factor was reduced. This is promising; perhaps TAO could be more equivalent to Sloan letters than shapes that share the Sloan dimensions. However, this outcome will change if the protocol differs (chart vs QUEST, for example), and will change when compared to other reference sets (Landolt C vs ETDRS, for example). In this sense, scaling will be an important feature of the evolution of the new TAO set. Lea symbols have been rescaled on more than one occasion in relation to different references,[22] and in some cases are individually scaled to allow equal blurring (see appendix I in Cyert, 2010[28] for details). An appropriate scaling factor (ideally for TAO as a set) should be assessed in a large cohort with identical protocols and a considered reference optotype set.

Although scaling relates directly to the cut-off used for screening, the AUC can be compared regardless of specific acuity cut-offs. TAO tests were highly effective at identifying children with vision impairment ($AUC_{TAOregular}$ = 0.96, $AUC_{TAOvanishing}$ = 0.95), but Parr was less effective (AUC_{Parr} = 0.82). In children aged 5–6 years there was an enhanced advantage of TAO ($AUC_{TAOregular}$ = 0.97, $AUC_{TAOvanishing}$ = 0.98) over Parr (AUC_{Parr} = 0.75). TAO results for sensitivity ($TAO_{regular}$ = 91%, TAO vanishing = 91%) and specificity ($TAO_{regular}$ = 85%, $TAO_{vanishing}$ = 82%) are in line with the electronic Jaeb Visual Acuity Screener, reporting optimal sensitivity and specificity of 90% and 83%, respectively.[32] As with LoA, trials used within the QUEST protocol can be vastly truncated if screening is the goal. Two to three trials for TAO regular (5–6 for vanishing) is sufficient to achieve an AUC of 0.90, after which additional trials achieve diminishing returns. Note also the large jump in screening accuracy between presenting a single trial and two trials. This is perhaps intuitive, as VA must be better than 0.2 logMAR to pass our

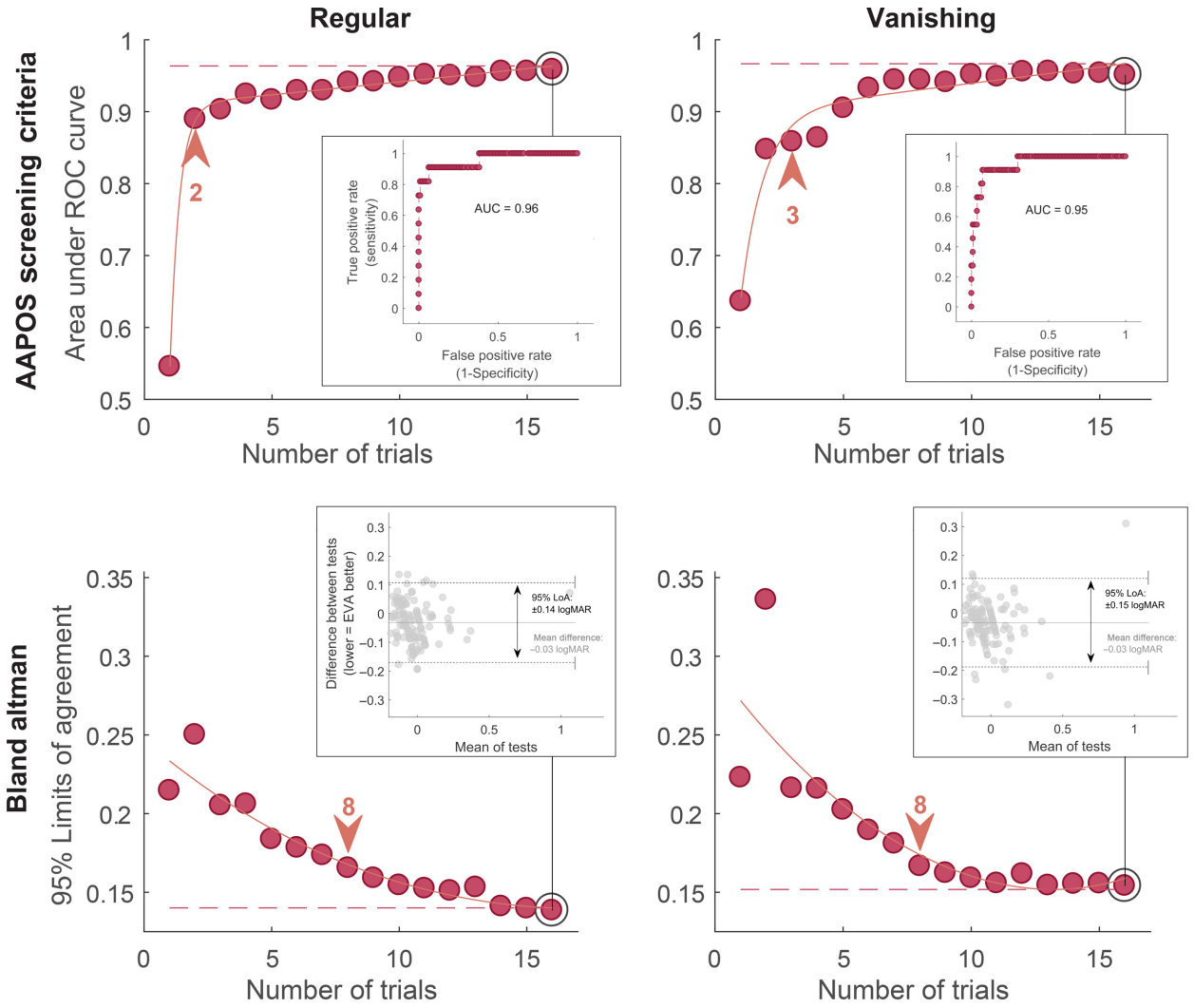

Figure 9. Impact of number of trials on test-efficacy using The Auckland Optotypes. The upper row shows area under receiver operating characteristic (ROC) curves for AAPOS screening criteria plotted against the number of trials administered. The lower row displays the results from Bland Altman 95% limit of agreement with Electronic Visual Acuity (EVA) with differing number of trials. Orange arrows indicate the trial closest to the knee point.

screening protocol. In the tablet protocol the first presentation was forced to 0.3 logMAR (a fail), and the second to 0.15 logMAR (a pass). Additional trials appear to help address lapses, correct guesses and to confirm a threshold between 0.15 and 0.3 logMAR. This confirmation with additional trials appears more important for the vanishing format. Further work on the relative benefits of repeat testing at threshold compared to the same number of measures at continuous stimulus levels around the threshold would be of interest.

For both LoA and ROC results, protocol was an important differentiating factor among tests. We were interested in whether a Bayesian adaptive staircase would be feasible/useful for children, having adjusted our implementation of this procedure based on our experience.[25] All children were testable and their lapse rates were comparable to adults,[20] suggesting the current protocol is feasible. Adding an automated matching phase (similar to a familiarisation phase with the EVA[1]) reduced the incidence of lapsing on the first trial compared to our previous project. The addition of a re-randomise function also allowed potential lapses to be averted by the assessor. This raises an important question about subjective aspects of childhood VA testing. Giving the assessor the ability to make a subjective decision about potential lapses (for example 'did you mean to say heart?' or providing an alternative at the same stimulus level) may prevent lapses, but also shifts the task away from a pure 10AFC, and potentially biases

results.[16] To avoid differential subjective input, several groups have attempted to completely automate the process. For example, Aslam *et al.*[35] designed a near acuity test for children run without an assessor. For a variety of reasons (likely including lapsing), test-retest LoA were high (±0.27 logMAR). The degree to which assessor feedback, enthusiasm, and rapport impact upon test results is not fully understood. Whether removing this source of variability through automation makes results more or less accurate is an important area for future work.

More work around use of the optotypes in individuals with varying ages and cognitive capacities would also be of value. Children (even 5 and 6 year olds) were capable of completing the test with 10 options. Indeed, this test was more effective at detecting vision impairment than the Parr test, which presented fewer options. However, the analysis of correct responses found that some shapes (e.g. ⌂ and ☪) were more popular than others, suggesting that there is potentially value in a truncated subset of TAO.

There was little difference between regular and vanishing optotypes. Practically, vanishing optotypes displayed on an electronic device are more difficult to control. For example, power-saving features of our testing device meant that temporary detachment from a power supply could lead to changes in the operation of the tablet display that interfered with perceptual vanishing. Although we worked around this (by having our software check the device was plugged in) such issues must be identified and managed to ensure tests will work as intended in community settings. Testing a cohort of children with more diverse visual conditions, including those with a wider range of unaided VA, amblyopia and ocular pathology, would be extremely valuable to determine whether there is a difference between regular and vanishing results in particular conditions, as is the case in some adult conditions.[11]

Together, agreement with EVA and capacity to differentiate between children with and without a visual anomaly suggest TAO symbols and the Bayesian adaptive staircases have value for testing children, particularly in comparison to the Parr tests, the current New Zealand screening tool. Further investigation of procedures including optimal scaling of TAO optotypes, efficient step sizes and use of automation would be useful, as would the testing of children with more severe visual impairment.

Acknowledgements

The work was supported by Cure Kids (Grant # 3562/3709768) and by the Robert Leitl Trust. We would like to thank Jay South, Myra Leung and Janice Yeoman for their help with data collection as well as Jing Chen, Leah Lawrence and Stephanie Wallen for their contribution to data collection and analysis as part of each of their undergraduate honours dissertations.

Conflicts of interest

The authors have no conflicts of interest and have no proprietary interest in any of the materials mentioned in this article. The Auckland Optotypes were developed by some of the authors of this paper, but these materials have been made freely available.

References

1. Holmes JM, Beck RW, Repka MX *et al.* The amblyopia treatment study visual acuity testing protocol. *Arch Ophthalmol* 2001; 119: 1345–1353.
2. Moke PS, Turpin AH, Beck RW *et al.* Computerized method of visual acuity testing: adaptation of the Amblyopia Treatment Study visual acuity testing protocol. *Am J Ophthalmol* 2001; 132: 903–909.
3. Beck RW, Moke PS, Turpin AH *et al.* A computerized method of visual acuity testing: adaptation of the early treatment of diabetic retinopathy study testing protocol. *Am J Ophthalmol* 2003; 135: 194–205.
4. Guo CX, Babu RJ, Black JM *et al.* Binocular treatment of amblyopia using videogames (BRAVO): study protocol for a randomised controlled trial. *Trials* 2016; 17: 504.
5. Cotter SA, Cyert LA, Miller JM *et al.* Vision screening for children 36 to G72 Months: recommended practices. *Optom Vis Sci* 2015; 92: 6–16.
6. Carkeet A. Modeling logMAR visual acuity scores: effects of termination rules and alternative forced-choice options. *Optom Vis Sci* 2001; 78: 529–538.
7. Hamm LM, Yeoman JP, Anstice N & Dakin SC. The Auckland Optotypes: an open-access pictogram set for measuring recognition acuity. *J Vis* 2018; 18: 13.
8. Frisen L. Vanishing optotypes. New type of acuity test letters. *Arch Ophthalmol* 1986; 104: 1194–1198.
9. Adoh TO, Woodhouse JM & Oduwaiye KA. The Cardiff test: a new visual acuity test for toddlers and children with intellectual impairment. A preliminary report. *Optom Vis Sci* 1992; 69: 427–432.
10. Shah N, Dakin SC, Redmond T & Anderson RS. Vanishing Optotype acuity: repeatability and effect of the number of alternatives. *Ophthalmic Physiol Opt* 2011; 31: 17–22.
11. Shah N, Anderson R, Tufail A, Egan C & Dakin S. Visual acuity loss in patients with AMD, measured using a vanishing optotype letter chart. *Invest Ophthalmol Vis Sci* 2013; 54: 5021.
12. Sloan LL. Measurement of visual acuity: a critical review. *AMA Arch Ophthalmol* 1951; 45: 704–725.
13. Bailey IL & Lovie JE. New design principles for visual acuity letter charts. *Am J Optom Physiol Opt* 1976; 53: 740–745.

14. Ferris III FL, Kassoff A, Bresnick GH & Bailey I. New visual acuity charts for clinical research. *Am J Ophthalmol* 1982; 94: 91–96.
15. Bailey IL & Lovie-Kitchin JE. Visual acuity testing. From the laboratory to the clinic. *Vision Res* 2013; 90: 2–9.
16. Bailey IL & Jackson AJ. Changes in the clinical measurement of visual acuity. *J Phys Conf Ser* 2016; 772: 012046.
17. Black JM, Jacobs RJ, Phillips G *et al.* An assessment of the iPad as a testing platform for distance visual acuity in adults. *BMJ Open* 2013; 3: pii: e002730.
18. Watson AB & Pelli DG. QUEST: a Bayesian adaptive psychometric method. *Percept Psychophys* 1983; 33: 113–120.
19. King-Smith PE, Grigsby SS, Vingrys AJ, Benes SC & Supowit A. Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation. *Vision Res* 1994; 34: 885–912.
20. Klein SA. Measuring estimating, and understanding the psychometric function: a commentary. *Percept Psychophys* 2001; 63): 1421–1455.
21. Toole AJ, Raasch TW, Fogt N & Brunstetter TJ. Evaluation of single pixel step sizes in visual acuity assessment. *Optom Vis Sci* 2011; 88: 244–250.
22. Hyvarinen L, Nasanen R & Laurinen P. New visual acuity test for pre-school children. *Acta Ophthalmol* 1980; 58: 507–511.
23. Milling A, Newsham D, Tidbury L, O'Connor A & Kay H. The redevelopment of the Kay picture test of visual acuity. *Br Ir Orthopt J* 2016; 13: 14–21.
24. Bland JM & Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1: 307–310.
25. Hamm LM, Langridge F, Yeoman J *et al. Childhood vision screening in Tonga*. International Congress of Paediatrics, Vancouver, Canada 2016.
26. Brainard DH. The Psychophysics Toolbox. *Spat Vis* 1997; 10: 433–436.
27. Kleiner M, Brainard D & Pelli D. What's new in Psychtoolbox-3. *Perception* 2007; 36: 1–16.
28. Vision In Preschoolers Study Group. Effect of age using Lea symbols or HOTV for preschool vision screening. *Optom Vis Sci* 2010; 87: 87–95.
29. Cromelin CH, Candy TR, Lynn MJ, Harrington CL & Hutchinson AK. The handy eye chart: a new visual acuity test for use in children. *Ophthalmology* 2012; 119: 2009–2013.
30. Prins N & Kingdom FAA. *Palamedes: Matlab routines for analyzing psychophysical data*, 2009. http://www.palamedestoolbox.org
31. Donahue SP, Arthur B, Neely DE, Arnold RW, Silbert D & Ruben JB. Guidelines for automated preschool vision screening: a 10-year, evidence-based update. *J AAPOS* 2013; 17: 4–8.
32. Yamada T, Hatt SR, Leske DA *et al.* A new computer-based pediatric vision-screening test. *J AAPOS* 2015; 19: 157–162.
33. Moganeswari D, Thomas J, Srinivasan K & Jacob GP. Test re-test reliability and validity of different visual acuity and stereoacuity charts used in preschool children. *J Clin Diagn Res* 2015; 9: NC01-5.
34. Ma DJ, Yang HK & Hwang JM. Reliability and validity of an automated computerized visual acuity and stereoacuity test in children using an interactive video game. *Am J Ophthalmol* 2013; 156:195–201.e1.
35. Aslam TM, Tahir HJ, Parry NRA *et al.* Automated measurement of visual acuity in pediatric ophthalmic patients using principles of game design and tablet computers. *Am J Ophthalmol* 2016; 170: 223–227.